

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 Penelitian Terkait**

Berikut beberapa penelitian yang meneliti terkait dalam Datamining :

1. Liliana Swastina pada tahun 2013. Penelitian ini membahas implementasi teknik datamining pada proses penentuan jurusan mahasiswa dengan menerapkan algoritma decision tree C4.5. uji coba di lakukan dengan menggunakan data yang di dapatkan dari BAA STMIK Indonesia Banjarmasin. Hasil menunjukan bahwa proses pembentukan pohon pada algoritma C4.5 akurat di terapkan untuk penentuan kesesuaian jurusan mahasiswa dengan tingkat keakuratan 93,31% dan akurasi rekomendasi jurusan sebesar 82,64%.
2. Faid Ari Prastya pada tahun 2015. Penelitian ini membahas implementasi datamining untuk memprediksi jurusan siswa SMA N 3 Rembang menggunakan algoritma C4.5. Data yang digunakan dalam penelitian ini adalah sumber data primer. Data di peroleh dari SMA N 3 Rembang. Data yang dikumpulkan adalah data siswa baru tahun ajaran 2014/2015. Akurasi yang dihasilkan dari datashet Siswa Baru SMA N 3 Rembang Tahun ajaran 2014/2015 pada rapidminer sebesar 74,65%.
3. Aradea,Satriyo A dan kawan-kawan pada tahun 2011 . penelitian ini membahas implementasi datamining untuk penentuan pola data penerimaan mahasiswa baru menggunakan gabungan dua algoritma antara C4.5 dan ID3. Data yang digunakan merupakan data calon mahasiswa yang mendaftar pada suatu program studi di suatu Universitas. Akurasi hasil dari study kasus yang dibahas mencapai nilai keakuratan yang maksimal yaitu 100%.
4. Kikie Rieski Andini dan kawan-kawan. Penelitian ini membahas mengenai Penerapan Datamining Untuk Mengolah Informasi Konsentrasi Keahlian Dengan Metode Clustering Pada Universitas Bina

Darma. Metode pengumpulan data dalam penelitian ini dilakukan dengan cara metode observasi, wawancara dan metode studi pustaka yang dinilai objektif dan relevan dengan inti permasalahan sebagai indikator keberhasilan.

5. Obbie Kristanto pada tahun 2014. Penelitian ini membahas tentang penerapan algoritma klasifikasi datamining ID3 untuk menentukan penjurusan siswa SMAN 6 Semarang. Pada penelitian ini dibuat suatu aplikasi system pengambilan keputusan untuk menentukan jurusan pada siswa SMAN 6 Semarang. Tingkat akurasi dengan menggunakan data dari guru BP sebagai perbandingan menghasilkan akurasi sebesar 80%.
6. Diana Laily Fithri dan Eko Darmanto pada tahun 2014. Penelitian ini membahas tentang system pendukung keputusan untuk memprediksi kelulusan mahasiswa menggunakan metode naïve bayes. Penelitian menggunakan data mahasiswa progdi system informasi Fakultas Teknik Universitas Muria Kudus. Penerapan algoritma naïve bayes menggunakan 8 parameter data mahasiswa menghasilkan tingkat akurasi 93%.
7. Yogi Yusuf W pada tahun 2007. Penelitian yang dibahas mengenai perbandingan performansi algoritma decision tree C5.0, CART dan CHAID untuk kasus prediksi status resiko kredit di suatu bank. Pada penelitian ini algoritma C5.0 menghasilkan rata-rata tingkat keakuratan sebesar 87,72% dengan standar deviasi 1,56, cart 87,28% standar deviasi 1,51 dan chaid 87,15% standar deviasi 2,19.
8. Kurniati Putri Wirdhaningsih dan kawan-kawan meneliti tentang penerapan algoritma decision tree C5.0 untuk peramalan forex. Parameter masukan data yang digunakan dalam prediksi adalah harga buka, harga terendah, harga tertinggi dan harga penutupan, sedangkan indicator yang digunakan prediksi naik turunnya nilai pertukaran adalah MACD dan RSI. Dalam penelitian penerapan algoritma decision tree C5.0 untuk peramalan forex secara optimal diperoleh dari time frame

M15 (per-15-menit) untuk pair mata uang EUR/USD dengan akurasi buy 84,49% dan akurasi sell 83,69% profit optimal sebesar 2,31\$.

9. Holisatul munawaroh dan kawan-kawan meneliti tentang perbandingan algoritma ID3 dan C5.0 dalam identifikasi penjurusan siswa SMA. Data yang digunakan dalam penelitian ini merupakan data siswa kelas X angkatan 2011/2012 yang diperoleh dari SMAN 2 Bangkalan sebanyak 200 data. Penelitian ini menghasilkan akurasi C5.0 95% lebih baik dari ID3 sebesar 93%.
10. Sofi Defiyanti dan D. L. Crispina Pardede meneliti tentang “perbandingan kinerja algoritma ID3 dan C4.5 dalam klasifikasi spam-mail. Pada penelitian tersebut kinerja algoritma ID3 lebih baik di banding algoritma C4.5.

Tabel 2. 1 Penelitian Terkait

No	Nama Peneliti	Tahun	Masalah	Algoritma	Hasil akurasi
1	Liliana Swastina	2013	Menentukan jurusan mahasiswa	Decision Tree C4.5	93,31%
2	Faid Ari Prastya	2015	Memprediksi jurusan siswa SMAN 3 Rembang	Decision Tree C4.5	Dengan Rapidminer akurasi 74,65%
3	Aardea, Satrio A, Ariyan Z, Yuliana A.	2011	Menentukan pola data penerimaan mahasiswa baru	Gabungan C4.5 dengan ID3	100%

4	Kikie Rieski Andini, M.Akbar, Helda Yudiasuti	-	Mengolah Informasi Konsentrasi Keahlian Pada Universitas Bina Darma	K-Means Clustering	Mengambil hasil akurasi tertinggi
5	Obbie Kristanto	2014	Menentukan penjurusan siswa SMAN 6 Semarang	ID3	80%

## 2.2 Pemilihan Jurusan Mahasiswa Teknik Informatika Universitas Dian Nuswantoro

Teknik Informatika merupakan salah satu jurusan pada Fakultas Ilmu Komputer di Universitas Dian Nuswantoro. Dua tahun terakhir mulai tahun 2012, Universitas Dian Nuswantoro menerapkan kurikulum baru, yakni membagi Teknik Informatika menjadi 2 bidang yaitu Informatika murni dan Rekayasa Perangkat Lunak. Kedua bidang tersebut wajib di pilih untuk di dalam oleh Mahasiswa Teknik Informatika yang akan menginjak semester ke-lima. Dua tahun pertama di terapkan kurikulum tersebut, mahasiswa masih banyak yang labil akan memilih bidang yang mereka tekuni termasuk penulis. Sebagian besar mahasiswa memilih informatika murni karena kabarnya yang lebih mudah dalam menempuh tugas akhir dan mata kuliah yang tidak terlalu sulit. Sebagai solusi kebingungan yang di alami mahasiswa, dilakukan penelitian untuk membuat suatu sistem yang dapat membantu mahasiswa dalam memilih bidangnya. Diharapkan sistem yang akan di buat, bermanfaat dan dapat membantu mahasiswa dalam memilih bidang sesuai bakat mereka.

## 2.3 Data Mining

### 2.3.1 Pengertian Data Mining

Beberapa ahli lain berpendapat, Datamining merupakan serangkaian proses yang digunakan untuk menggali nilai tambah dari kumpulan data yang berupa pengetahuan dan tidak diketahui secara

No	Nama Peneliti	Tahun	Masalah	Algoritma	Hasil akurasi
6	Diana Laily Fithri dan Eko Darmanto	2014	Menentukan keputusan untuk memprediksi kelulusan mahasiswa	Naïve Bayes	93%
7	Yogi Yusuf W	2007	membandingkan performansi algoritma decision tree C5.0, CART dan CHAID untuk kasus prediksi status resiko kredit di suatu bank	Decision Tree C5.0, CART, CHAID	C5.0 87,72%, CART 87,28%, CHAID 87,15%
8	Kurniati Putri Wirdhaningsih dan kawan-kawan	-	penerapan algoritma decision tree C5.0 untuk peramalan forex	Decision Tree C5.0	akurasi buy 84,49% dan akurasi sell 83,69%
9	Holisatul munawaroh dan kawan-kawan	-	membandingkan algoritma ID3 dan C5.0 dalam identifikasi penjurusan siswa SMA	perbandingan C5.0 dengan ID3	akurasi C5.0 95% lebih baik dari ID3 sebesar 93%
10	l. Sofi Defiyanti dan D. L. Crispina Pardede	-	meneliti tentang “perbandingan kinerja algoritma ID3 dan C4.5 dalam klasifikasi spam-mail”	Perbandingan ID3 dan C4.5	kinerja ID3 lebih baik di banding C4.5

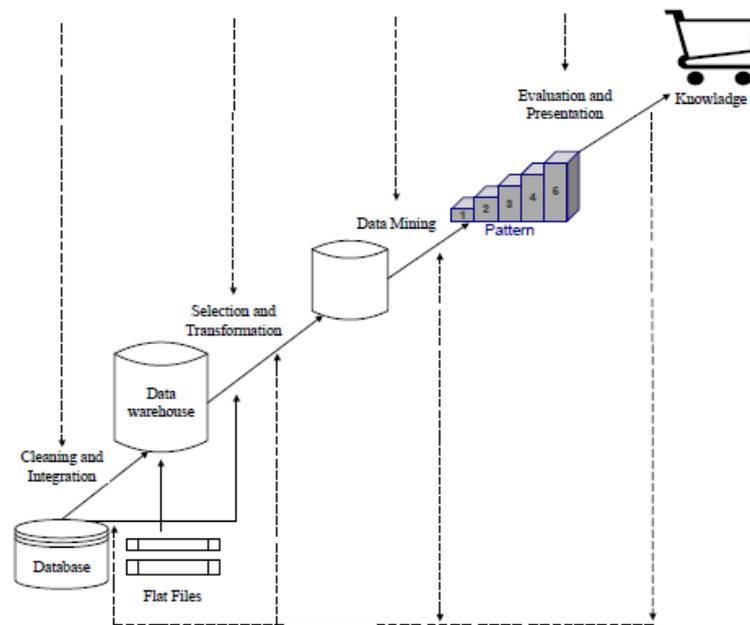
manual. Istilah Data mining dengan *knowledge discovery in databases* atau biasa disingkat dengan KDD kerap digunakan secara bergantian untuk menjelaskan proses penggalian informasi yang tersembunyi dalam suatu basis data yang besar. Kedua konsep tersebut memiliki konsep yang berbeda, akan tetapi berkaitan antara satu dengan lain [14].

Beberapa pekerjaan yang berkaitan dengan datamining dapat di bagi menjadi empat kelompok yaitu model prediksi, analisis cluster, analisis asosiasi dan deteksi anomaly. Salah satu contohnya yaitu melakukan pendeteksian jenis penyakit pada pasien berdasarkan sejumlah nilai-nilai parameter penyakit yang di derita. Pekerjaan ini termasuk jenis klasifikasi. [16]

Dalam datamining selalu ada set data. Bukan datamining namanya jika tidak ada data set atau set data yang diolah di dalamnya. Kata '*data*' dalam terminology statistic adalah kumpulan objek dengan atribut-atribut tertentu, dimana objek tersebut adalah individu berupa data dimana setiap data memilih atribut. Atribut tersebut berpengaruh pada dimensi dari data, semakin banyak atribut atau fitur maka semakin besar dimensi data. Kumpulan data-data membentuk set data. Dalam buku ini [17] kadang menyebut data, kadang menyebut vektor, keduanya mempunyai maksud yang sama.

### 2.3.2 Tahap – Tahap Data Mining

Merupakan sebuah tuntutan bagi data mining adalah saat diterapkan pada suatu data berskala besar adalah diperlukannya metodologi sistematis. Datamining dikenal sebagai proses yang memiliki beberapa tahapan serta memiliki umpan balik dari setiap tahapan ke tahapan sebelumnya. Beberapa tahapan tersebut seperti pada gambar 2.1



Gambar 2. 1 Tahapan Datamining

Proses KDD secara garis besar dapat dijelaskan sebagai berikut [11]:

### 1. Seleksi Data

Pada tahap ini dilakukan penyeleksian data dimana tidak semua data akan dipakai. Pemilihan data harus dilakukan sebelum lanjut pada tahap penggalian informasi pada proses KDD. Data yang telah di seleksi disimpan pada suatu basis data yang terpisah.

### 2. Pembersihan Data

Proses pembersihan pada data merupakan focus KDD, yang dilaksanakan sebelum dilakukan proses datamining. Proses pembersihan antara lain pembuangan duplikasi data, pemeriksaan data inkonsisten dan memperbaiki kesalahan pada data maupun informasi eksternal. Proses yang lain yaitu proses memperkaya data lengkap yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD meliputi data dan informasi eksternal.

### 3. Transformasi Data

Pada tahap ini data dikonversikan ke dalam format yang sesuai agar mudah di proses dalam datamining.

### 4. Data mining

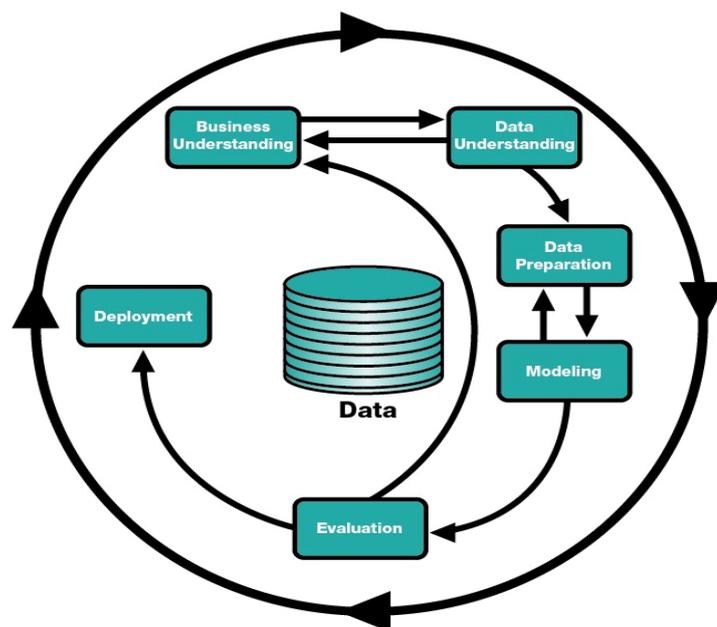
Proses ini merupakan proses utama saat metode diterapkan untuk menemukan pengetahuan yang penting ataupun data yang tersembunyi.

### 5. Interpretation / Evaluation

Tahap ini adalah bagian dari proses KDD yang meliputi pemeriksaan pola atau informasi yang ditemukan apakah bertentangan dengan fakta atau hipotesa yang sudah ada sebelumnya.

#### 2.3.3 CRISP – DM

CRISP-DM atau biasa di singkat CROSS-Industry Standard Process for Data Mining adalah proses standarisasi dalam datamining yang terbukti dapat diaplikasikan pada sebagian sektor industri. Dibawah ini merupakan gambar proses siklus pada hidup pengembangan CRISP-DM [8]:



Gambar 2. 2 CRISP-DM

**Penjelasan tahapan di atas adalah sebagai berikut :**

1. Business Understanding

Tahap pertama CRISP-DM adalah memahami tujuan serta kebutuhan sudut pandang bisnis, kemudian terjemahkan pengetahuan ini ke dalam pendefinisian masalah datamining. Setelah itu tentukan rencana dan strategi untuk sampai ke tujuan yang di rencanakan.

2. Data Understanding

Pada tahap kedua dimulai pengumpulan data yang akan dilanjutkan dengan proses mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. Data Preparation

Tahap yang ketiga meliputi semua kegiatan pembangunan dataset akhir atau data yang akan diproses pada tahap modeling dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini mencakup pemilihan tabel, record, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data yang kemudian dijadikan masukan pada tahap pemodelan (modeling).

4. Modeling

Tahap ini dilakukan pemilihan dan penerapan berbagai teknik pemodelan sekaligus beberapa parameternya akan disesuaikan untuk mendapatkan nilai optimal. Ada beberapa teknik berbeda yang dapat diterapkan pada masalah data mining yang sama. Pada pihak lain terdapat teknik pemodelan yang membutuhkan format data khusus. Sehingga pada tahap ini masih dimungkinkan kembali ke tahap sebelumnya.

## 5. Evaluation

Pada tahap evaluasi, model telah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisa data. Pada tahap ini dilakukan evaluasi terhadap keefektifan dan kualitas model sebelum digunakan dan digunakan sebagai penentu apakah model dapat mencapai tujuan yang ditetapkan pada fase awal (Business Understanding). Kunci dari evaluasi adalah menentukan masalah bisnis yang belum dipertimbangkan. Akhir dari tahap ini harus ditentukan penggunaan hasil proses data mining.

## 6. Deployment

Pada tahap final, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap deployment dapat berupa pembuatan laporan maupun implementasi proses data mining yang berulang dalam perusahaan. Pada beberapa kasus, tahap deployment melibatkan konsumen di samping analisis data, karena sangat penting bagi konsumen untuk memahami tindakan yang harus dilakukan untuk menggunakan model yang telah dibuat.

### 2.3.4 Teknik Datamining

Proses data mining secara sistematis terdapat tiga langkah utama dalam data mining [19], diantaranya Eksplorasi atau pemrosesan awal data yang terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur dan lainnya. Proses yang kedua adalah membangun model dan melakukan validasi terhadapnya yaitu dengan melakukan analisis berbagai model dan memilih model dengan kinerja prediksi yang terbaik. Dalam langkah ini digunakan metode metode seperti klasifikasi, regresi, analisis cluster, deteksi anomaly dan lain sebagainya. Langkah ketiga atau terakhir adalah

penerapan, yang berarti di terapkannya suatu model pada data yang baru untuk menghasilkan prediksi masalah yang di investigasi.

Beberapa teknik dan sifat datamining [18] adalah sebagai berikut :

1. Classification (Predictive)
2. Clustering (Descriptive)
3. AssociationRule discovery (Descriptive)
4. SequentialPattern Discovery (Descriptive)
5. Regression (Predictive)
6. DeviationDetection (Predictive)

Namun teknik yang akan digunakan dalam penelitian ini yaitu teknik klasifikasi atau Classification.

### 2.3.5 Klasifikasi

Klasifikasi adalah [19] proses untuk menentukan sebuah record data baru pada salah satu dari beberapa kategori (klas) yang sudah di definisikan sebelumnya. Disebut juga '*supervised learning*'. Metode klasifikasi mengutamakan pengelompokan data dengan menerapkan salah satu algoritma. Metode klasifikasi ini berguna untuk proses bisnis yang membutuhkan informasi kategoris.

Klasifikasi juga termasuk metode analisis data yang digunakan sebagai pembentuk model yang memprediksi trend data. Data pada pohon keputusan sering dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan parameter sebagai kriteria dalam pembentukan pohon [8]. Berikut beberapa aplikasi dari klasifikasi:

1. Penjualan langsung (*Direct Marketing*)

Tujuan dari klasifikasi ini adalah mengurangi costsurat menyurat degan menentukan (targetin) satu set konsumen yang

mempunyai kesamaan dalam membeli produk telepon selular baru.

## 2. Fraud Detection

Tujuannya adalah memprediksi kasus-kasus transaksi curang dengan menggunakan kartu kredit.

## 3. Customer Attrition / Churn

Yaitu untuk memprediksi pelanggan mana yang akan berpindah ke kompetitor kita.

Metode–metode yang telah di kembangkan oleh periset untuk menyelesaikan kasus klasifikasi antara lain sebagai berikut [15] :

- a. Pohon Keputusan
- b. Pengklasifikasi Bayes / Naïve Bayes
- c. Jaringan Syaraf Tiruan
- d. Analisis Statistik
- e. Algoritma Genetik
- f. Rough Sets
- g. Pengklasifikasi K-nearest Neighbour
- h. Metode Berbasis Aturan

## 2.4 Decision Tree

Pohon keputusan adalah suatu metode klasifikasi dan prediksi yang dipercaya sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi sebuah pohon keputusan yang mempresentasikan aturan. Aturan tersebut dapat dipahami dengan mudah menggunakan bahasa alami. Mereka dapat di ekspresikan dalam bentuk bahasa basis data seperti *structured query language* untuk mencari *record* pada kategori tertentu.

Pohon keputusan berguna untuk mengeksplorasi data dan menemukan hubungan tersembunyi antara sejumlah variabel input dengan sebuah variabel target. Variabel tujuan sering kali dikelompokkan secara pasti dan model pada

pohon keputusan lebih mengarah pada perhitungan probabilitas dari tiap-tiap record terhadap kategori tersebut atau untuk mengklasifikasikan record dengan mengelompokkannya dalam satu kelas. Pohon keputusan dapat digunakan sebagai estimasi nilai dari variabel continue walaupun terdapat beberapa teknik yang lebih sesuai untuk kasus ini.

Banyak algoritma terpakai dalam pembentukan pohon keputusan seperti ID3, CART dan C4.5. Data yang ada dalam pohon keputusan biasanya dinyatakan ke bentuk tabel dengan atribut dan record. Atribut merupakan sebuah parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentuka main tenis, kriteria yang diperlukan adalah cuaca, angin dan temperature. Salah satu atriibut merupakan atribut yang menyatakan data solusi per item data yang disebut target atribut. Atribut memiliki nilai nilai yang dinamakan dengan instance [18].



Gambar 2. 3 Contoh pohon keputusan

## 2.5 Algoritma ID3

Algoritma *Iterative Dichotomiser 3* atau biasa disingkat dengan ID3 adalah salah satu algoritma yang digunakan pada metode Decision Tree. Algoritma Id3 memakai konsep informasi gain serta Entropy. Beberapa tahapan dalam kinerja ID3 dijelaskan seperti dibawah:

1. Untuk mencari informasi gain dapat menggunakan rumus :

$$Gain(S, A) = Entropy(S) - \sum_{v \in \text{nilai}(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Dimana :

- A : Atribut
- V : Suatu nilai yang mungkin atribut A
- Vnilai A : Himpunan yang mungkin untuk atribut A
- |Sv| : Jumlah sampel untuk nilai v
- |S| : Jumlah seluruh sampel data
- Entropy (Sv) : Entropy untuk sampel – sampel yang memiliki nilai v

Dengan rumus mencari Entropy sebelumnya :

$$Entropy(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

Keterangan :

- S : Ruang (data) sample yang digunakan untuk training.
- P+ : Jumlah yang bersolusi positif pada data sample untuk kriteria tertentu.
- P\_ : Jumlah yang beresolusi negative pada data sanple untuk riteria tertentu.

2. Setelah entropy diperoleh, langkah selanjutnya menentukan nilai gain terbesar.
3. Nilai gain terbesar akan membentuk simpul yang berisi atribut sebagai akar.
4. Proses perhitungan gain terus dilaksanakan sampai semua data telah termasuk dalam kelas yang sama. Atribut yang telah dipilih tidak diikuti lagi dalam perhitungan nilai *information gain*.

## 2.6 Confusion Matrix

Confusion Matrix adalah alat (tools) visualisasi yang biasa digunakan pada supervised learning. Tiap kolom pada matrix adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya.

Tabel 2.3 adalah contoh tabel confusion matrix yang menunjukkan klasifikasi dua kelas.

Tabel 2. 2 confusion matrix

		Prediksi	
		A	C
Aktual	Negatif	A	C
	Positif	B	D

Keterangan :

A = jumlah prediksi tepat bahwa instance bersifat negative

B = jumlah prediksi salah bahwa instance bersifat positif

C = jumlah prediksi salah bahwa instance bersifat negative

D = jumlah prediksi tepat bahwa instance bersifat positif

Beberapa persyaratan standar yang telah didefinisikan untuk matrix klasifikasi dua kelas :

- a. Keakuratan (AC) adalah proporsi jumlah prediksi benar. Rumus persamaannya adalah :

$$AC = A + D / A + B + C + D$$

- b. Syarat penarikan kembali (recall) atau tingkat positif benar (TP) yaitu proporsi kasus positif yang diidentifikasi dengan benar, yang di hitung dengan persamaan :

$$TP = D/C+D$$

- c. Tingkat positif salah (FP) yaitu proporsi kasus negatif yang salah diklasifikasikan sebagai positif, dihitung dengan menggunakan persamaan :

$$FP = B/A+B$$

- d. Tingkat negatif sejati (TN) didefinisikan sebagai proporsi kasus negatif yang diklasifikasikan dengan benar, dihitung dengan menggunakan persamaan :

$$TN = A/A+B$$

- e. Tingkat negatif palsu (FN) yaitu proporsi kasus positif yang salah diklasifikasikan sebagai negatif, dihitung dengan menggunakan persamaan :

$$FN = C/C+D$$

- f. Presisi (P) yaitu proporsi prediksi kasus positif yang benar, dihitung dengan menggunakan persamaan :

$$P = D/B+D$$

## 2.7 RapidMiner

RapidMiner adalah perangkat lunak yang bersifat terbuka (open source). RapidMiner merupakan solusi untuk melakukan suatu analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada

pengguna sehingga dapat membuat keputusan yang paling baik. Rapid Miner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. Rapid Miner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan bahasa java sehingga dapat bekerja pada semua sistem operasi.

Rapid Miner menyediakan GUI (Graphic User Interface) untuk merancang sebuah pipeline analitis. GUI ini akan menghasilkan file XML (Extensible Markup Language) yang mendefinisikan proses analitis keinginan pengguna untuk diterapkan ke data. File ini kemudian dibaca oleh Rapid Miner untuk menjalankan analisis secara otomatis.

Rapid Miner memiliki beberapa sifat sebagai berikut:

- a. Ditulis dengan bahasa pemrograman Java sehingga dapat dijalankan di berbagai sistem operasi.
- b. Proses penemuan pengetahuan dimodelkan sebagai operator trees.
- c. Representasi XML internal untuk memastikan format standar pertukaran data.
- d. Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
- e. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
- f. Memiliki GUI, command line mode, dan Java API yang dapat dipanggil dari program lain.

Beberapa Fitur dari Rapid Miner, antara lain:

- a. Banyaknya algoritma data mining, seperti decision tree dan self-organization map.
- b. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots.

- c. Banyaknya variasi plugin, seperti text plugin untuk melakukan analisis teks.
- d. Menyediakan prosedur data mining dan machine learning termasuk: ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi
- e. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI
- f. Mengintegrasikan proyek data mining Weka dan statistika R [8].

## 2.8 PHP

PHP singkatan dari Hypertext Preprocessor adalah bahasa skrip yang dapat ditanamkan atau disisipkan ke dalam HTML. PHP banyak dipakai untuk memrogram situs web dinamis. PHP dapat digunakan untuk membangun sebuah CMS. PHP di kembangkan pada tahun 1995 oleh Rasmus Lerdorf, dan sekarang dikelola oleh The PHP Group. PHP disebut bahasa pemrograman server side karena PHP diproses pada komputer server. Hal ini berbeda dibandingkan dengan bahasa pemrograman client-side seperti JavaScript yang diproses pada web browser (client).

Pada awalnya PHP merupakan singkatan dari *Personal Home Page*. Sesuai dengan namanya, PHP digunakan untuk membuat website pribadi. Dalam beberapa tahun perkembangannya, PHP menjelma menjadi bahasa pemrograman web yang powerful dan tidak hanya digunakan untuk membuat halaman web sederhana, tetapi juga website populer yang digunakan oleh jutaan orang seperti wikipedia, wordpress, joomla, dll. PHP dapat digunakan dengan gratis dan bersifat *Open Source*. PHP dirilis dalam lisensi *PHP License*, sedikit berbeda dengan lisensi *GNU General Public License (GPL)* yang biasa digunakan untuk proyek *Open Source*. Kemudahan dan kepopuleran PHP sudah menjadi standar bagi programmer web di seluruh dunia.

## 2.9 Kerangka Pemikiran

Berikut gambaran keseluruhan penelitian dalam bentuk diagram kerangka pemikiran :

Tabel 2. 3 Kerangka Pemikiran

<b>Masalah</b>		
Menerapkan Algoritma ID3 pada sistem prediksi minat studi Mahasiswa Teknik Informatika		
<b>Eksperimen/Analisa*</b>		
Tools	Data	Metode
RapidMiner dan PHP	Data Mahasiswa Teknik Informatika UDINUS angkatan 2012 – 2013.	Algoritma ID3 dan Waterfall
<b>Pengujian dan validasi hasil</b>		
Menguji metode dengan menggunakan data testing, serta menghitung precision dan recall dengan confusion matrix untuk mengetahui akurasi. Untuk pengujian sistem menggunakan Black Box Testing.		
<b>Hasil</b>		
Rekomendasi prediksi minat studi mahasiswa		